Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms

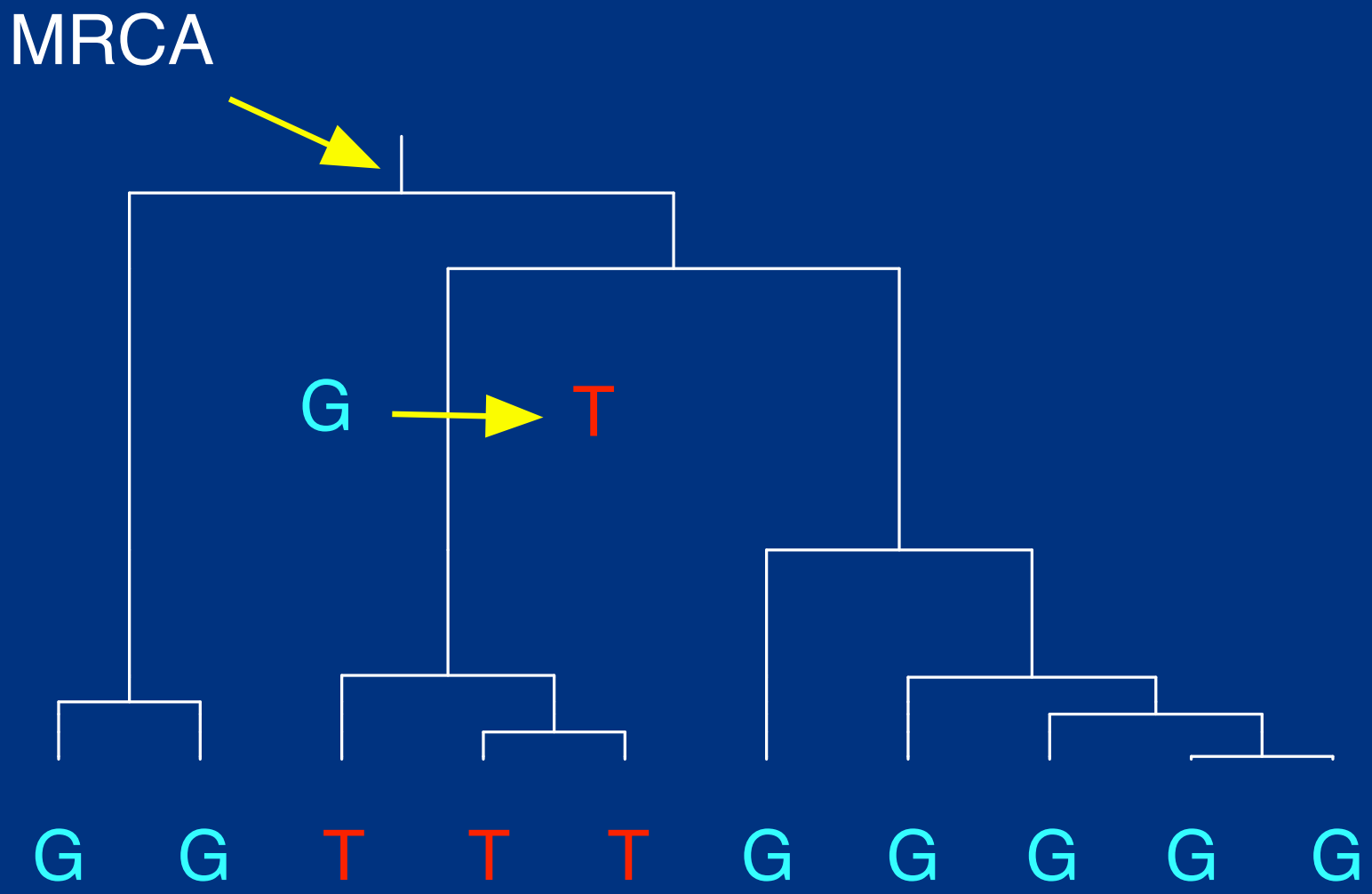
Magnus Nordborg

University of Southern California

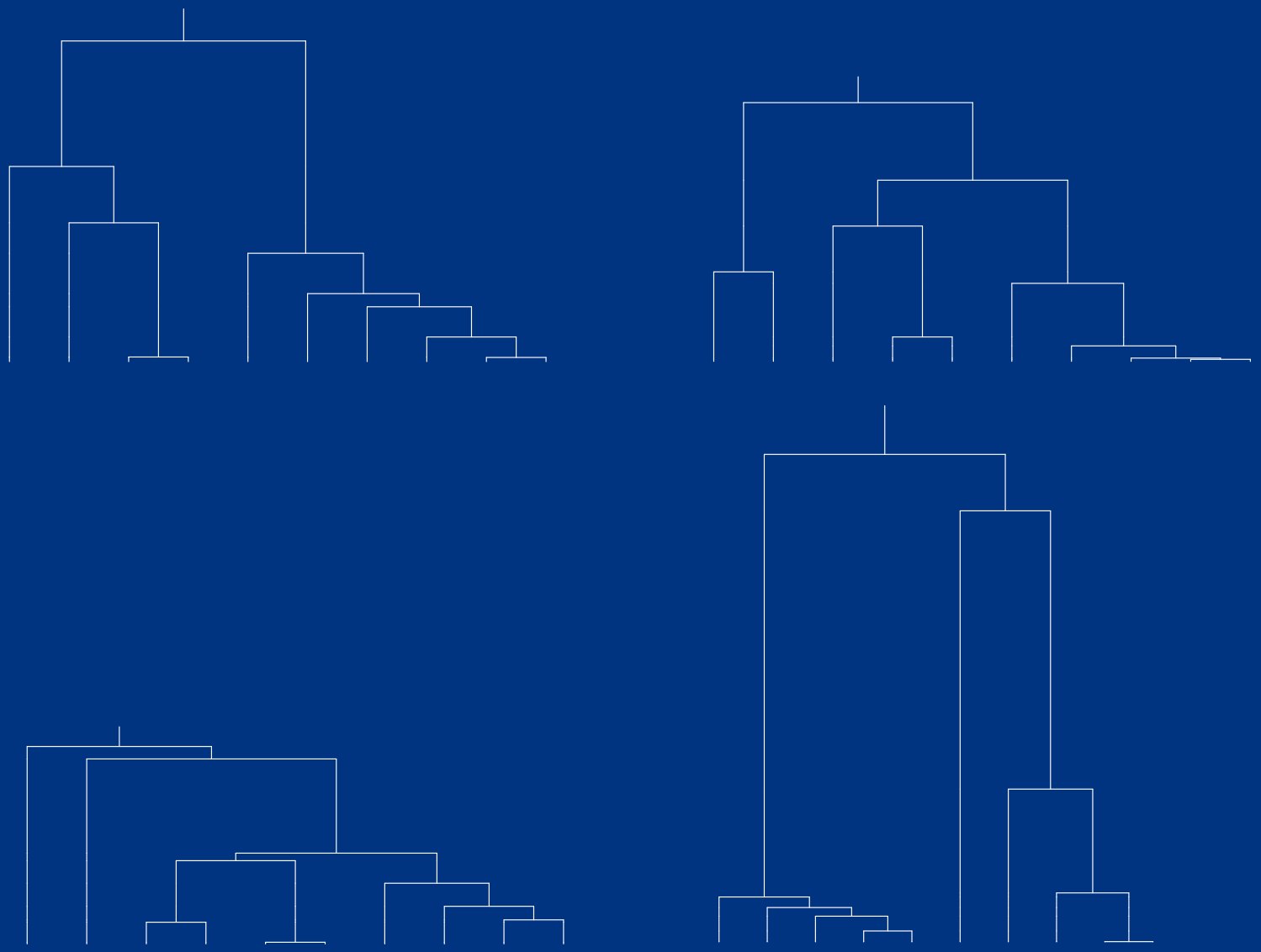
The importance of history

- Genetic polymorphism data represent the outcome of a single, highly complex, non-repeatable evolutionary history
- Traditional analysis methods cannot take this into account
- The stochastic process known as "the coalescent" presents a coherent statistical framework for analyzing genetic polymorphism data

The importance of history: mutations are random



The importance of history: trees are random

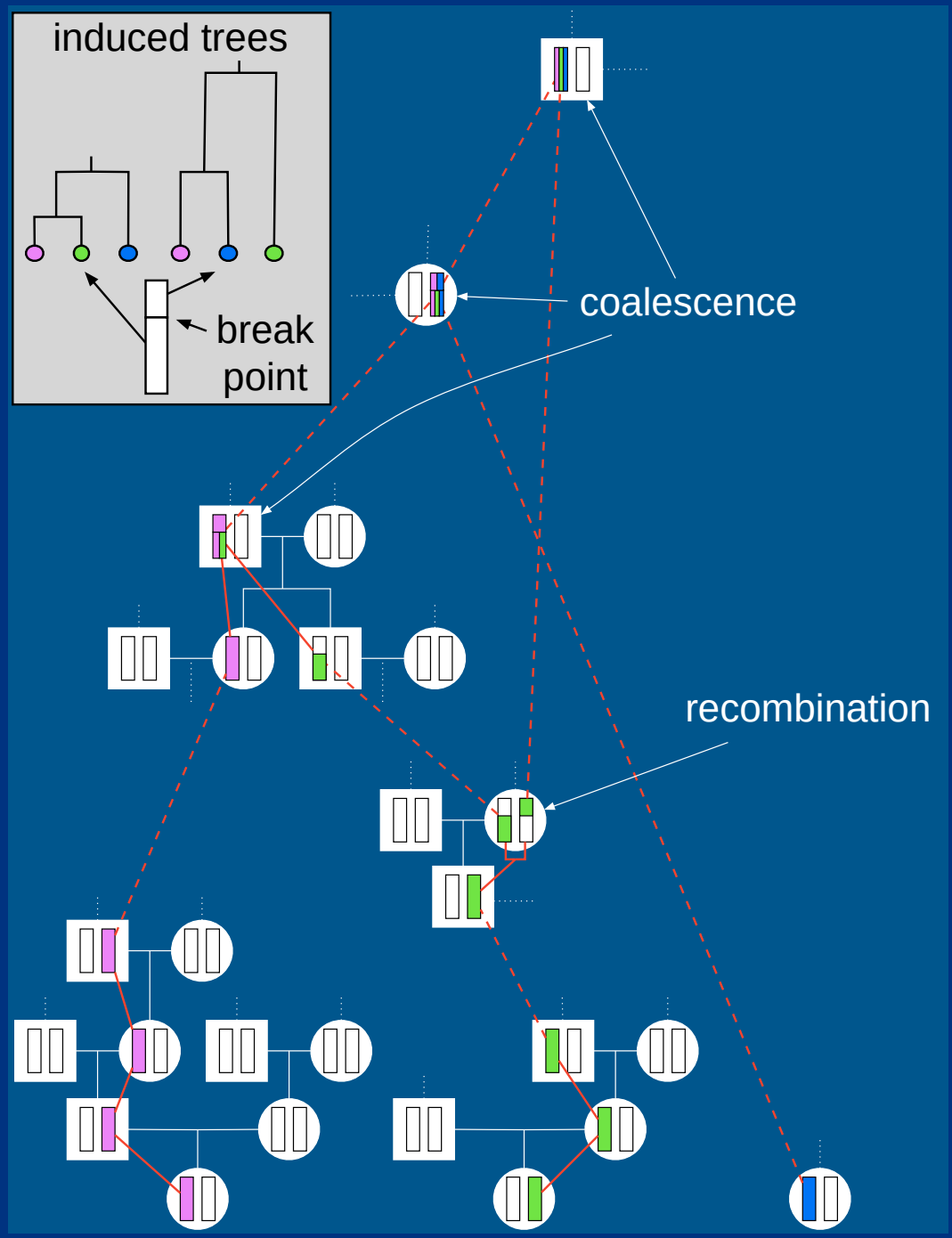


Modeling genetic polymorphism

At a minimum, models must include:

- coalescence (who begat whom, and when)
- mutation
- recombination

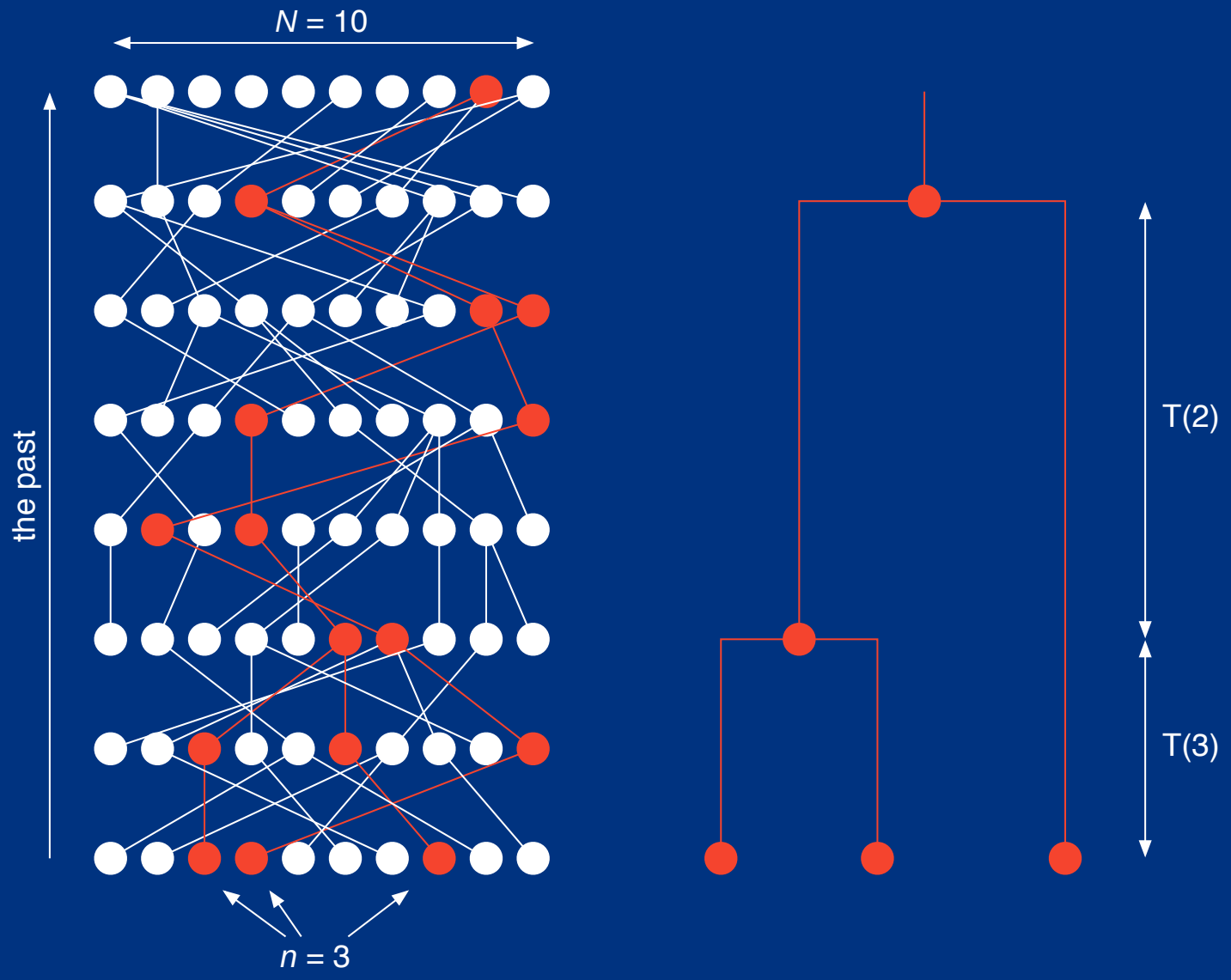
Recombination makes it possible for linked sites to have different genealogies



What is the coalescent?

- The coalescent is a stochastic process that is well-suited for modeling polymorphism data
- It is a natural extension to classical population genetics models

Coalescence: picking parents



The rate of coalescence

The rate at which lineages find each other depends on:

- The population size: the per-generation probability of coalescence is $\propto 1/N$
- The number of lineages: the rate of coalescence when there are k lineages is $\binom{k}{2}$
- A number of other demographic factors, such as inbreeding, age structure, and the variance in reproductive success

Because the per-generation probability of coalescence is on the order of $1/N$, we use a continuous-time approximation where time is measured in units of N generations

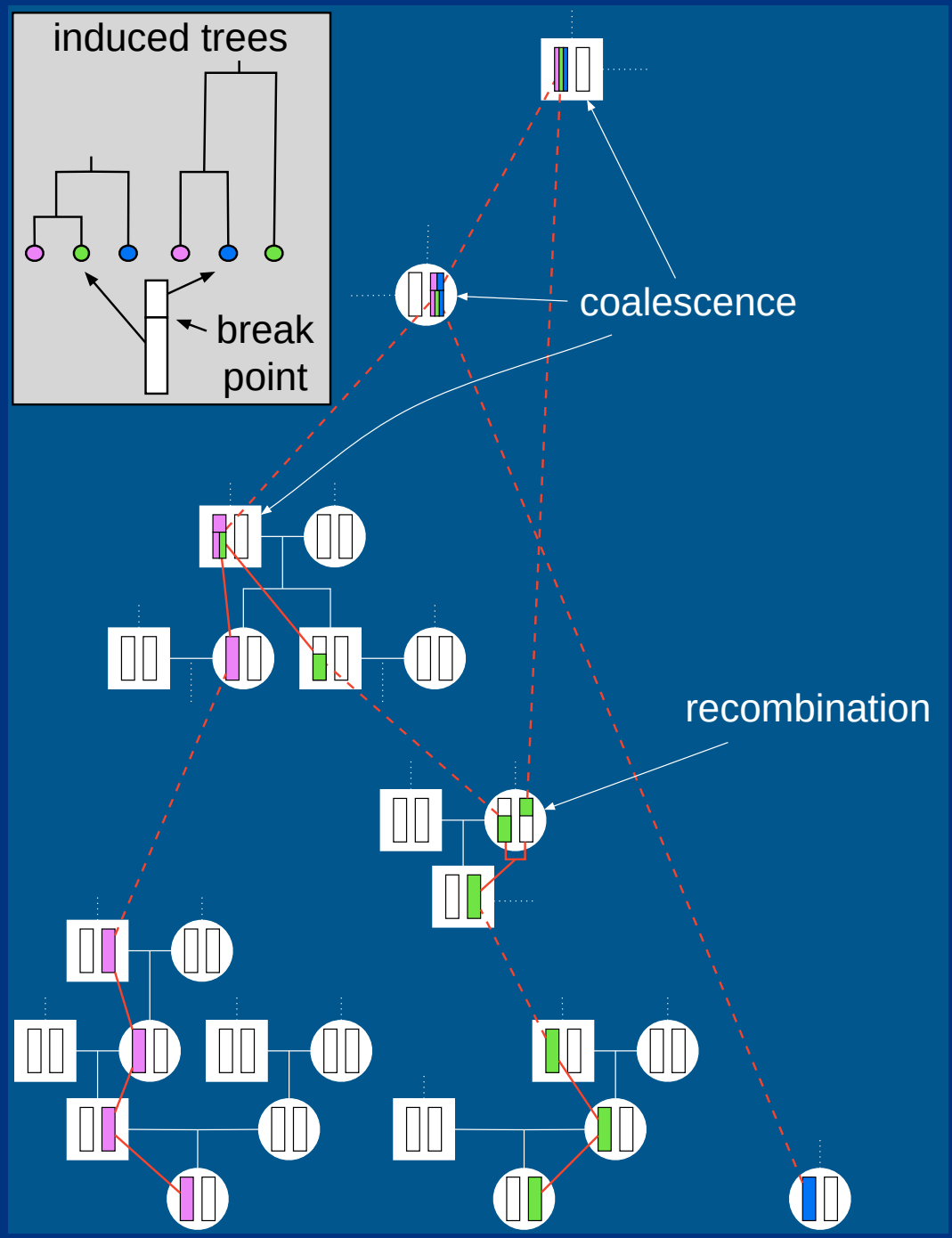
Mutation

- Selectively neutral mutations are added to the branches of the tree afterwards according to a rate that depends on the per-generation probability of mutation
- The expected number of mutations on a branch depends on its length — the expected number of mutations on the tree depends on the total branch length of the tree
- Any mutation model can be used

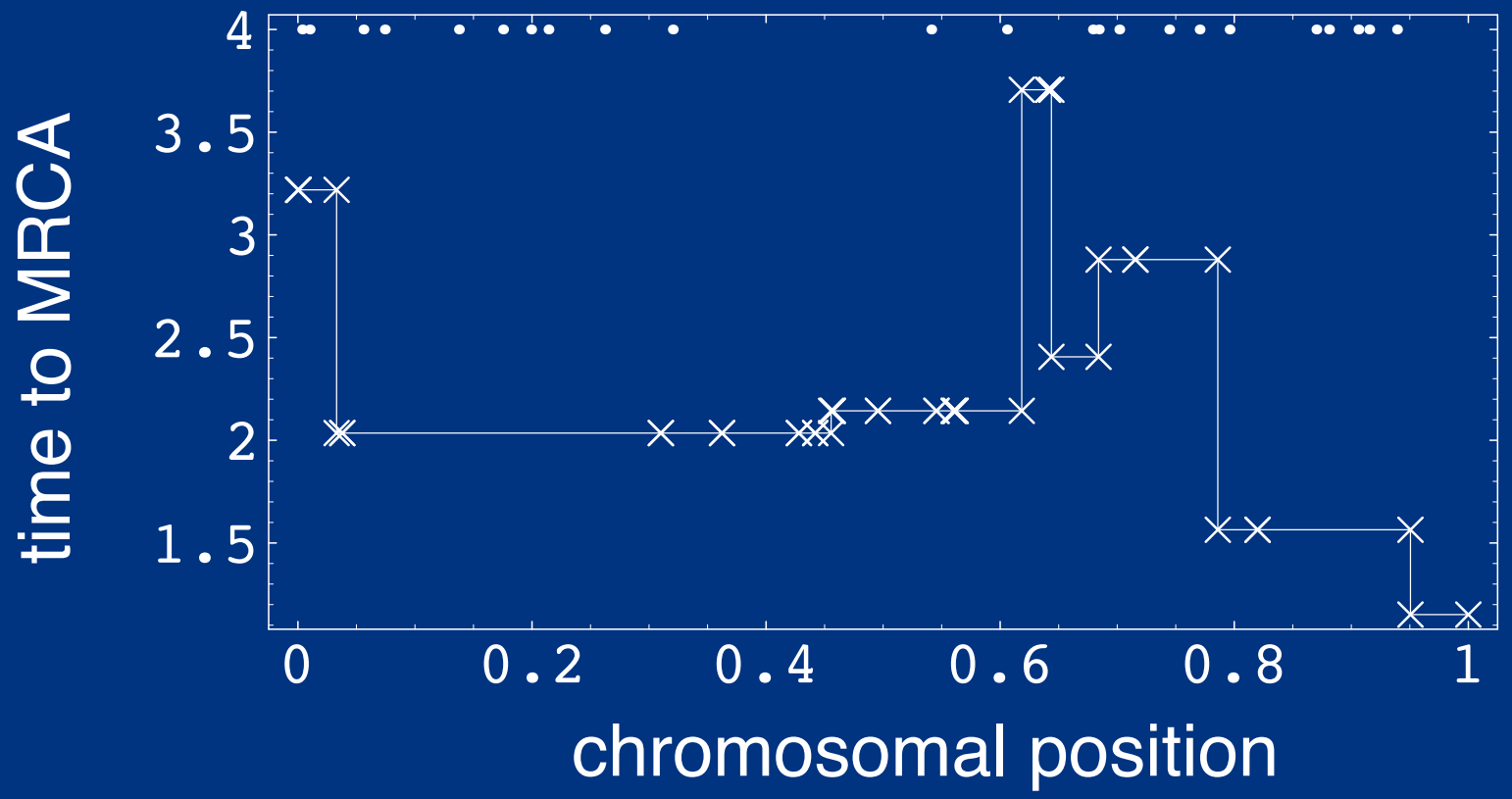
Recombination

- Recombination breaks up lineages according to a rate that depends on the per-generation probability of recombination
- There will be more recombination in the genealogy of a longer chromosomal segment
- Any recombination model can be used
- The coalescent with recombination generates a random graph — or a forest of trees

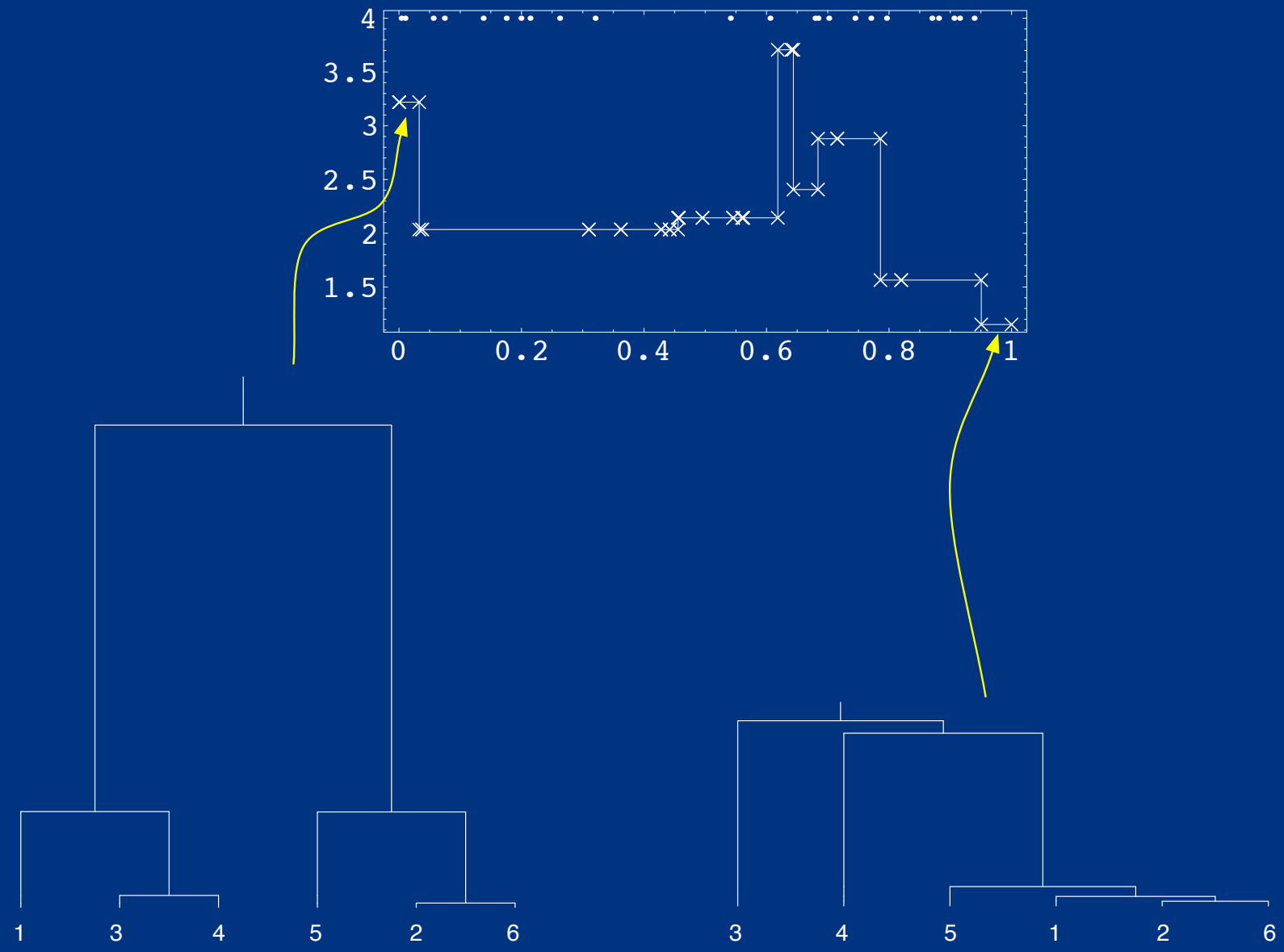
A graph or a forest. . .



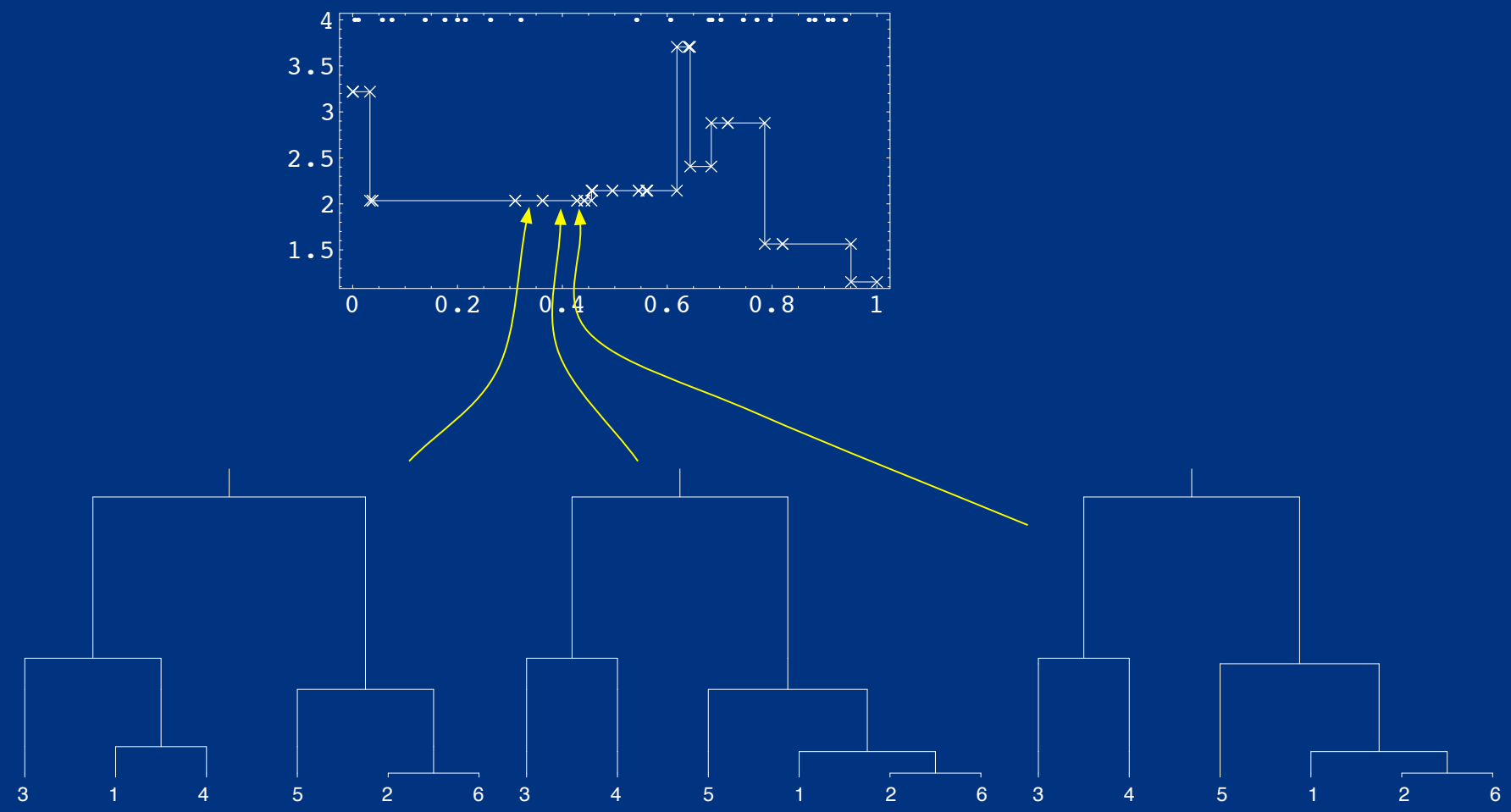
A walk through tree space



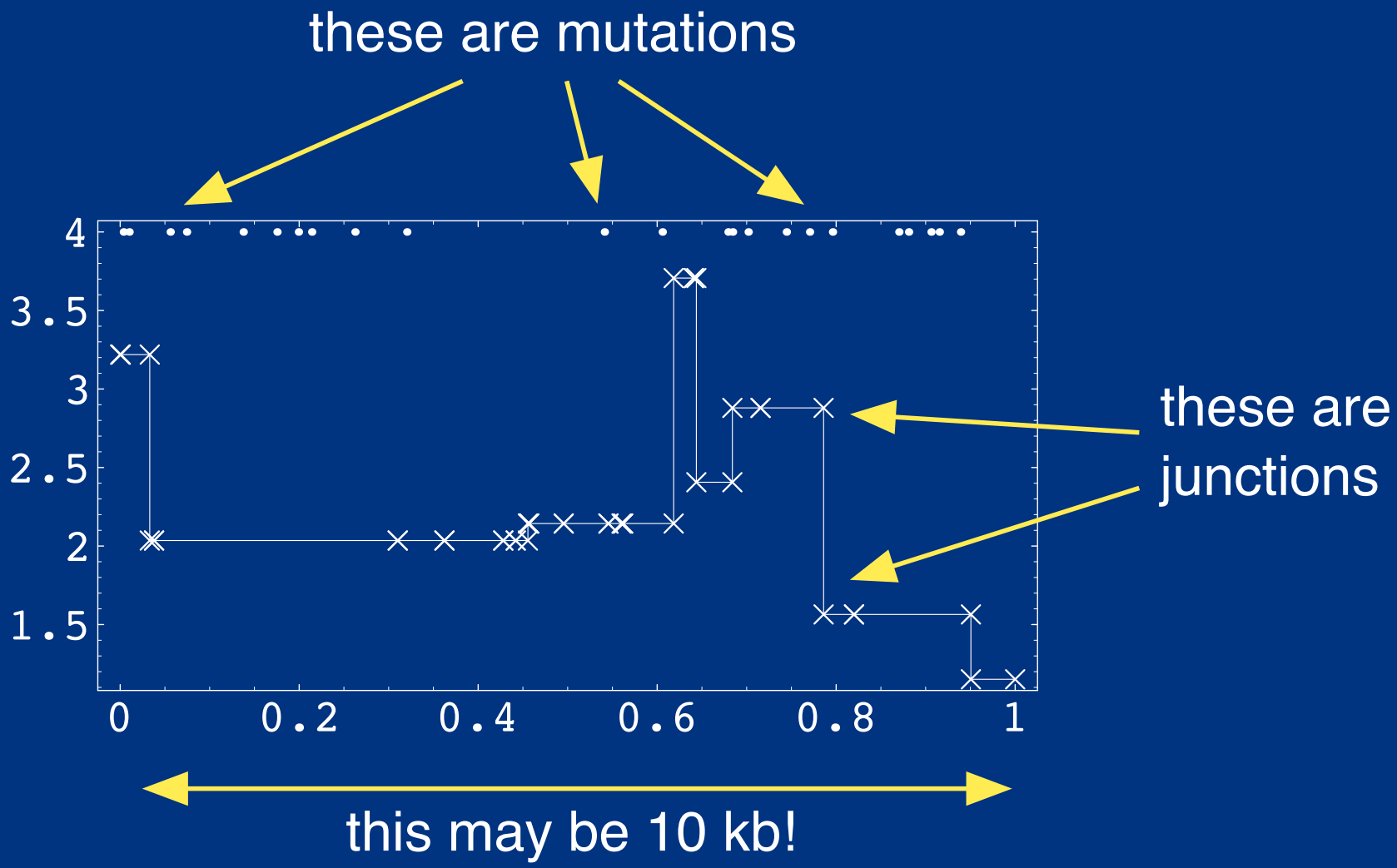
The trees are correlated



The trees are correlated



Recombination is common



Recombination is as common as mutation

- If 1 cM $\sim$ 1 Mb, then the probability of recombination per bp per generation is $\sim 10^{-8}$
- The probability of mutation per bp per generation is estimated to be *at most* 10^{-8}
- It follows that a sample of sequences will contain as many junctions as polymorphisms

Genealogical graphs can in general not be reconstructed

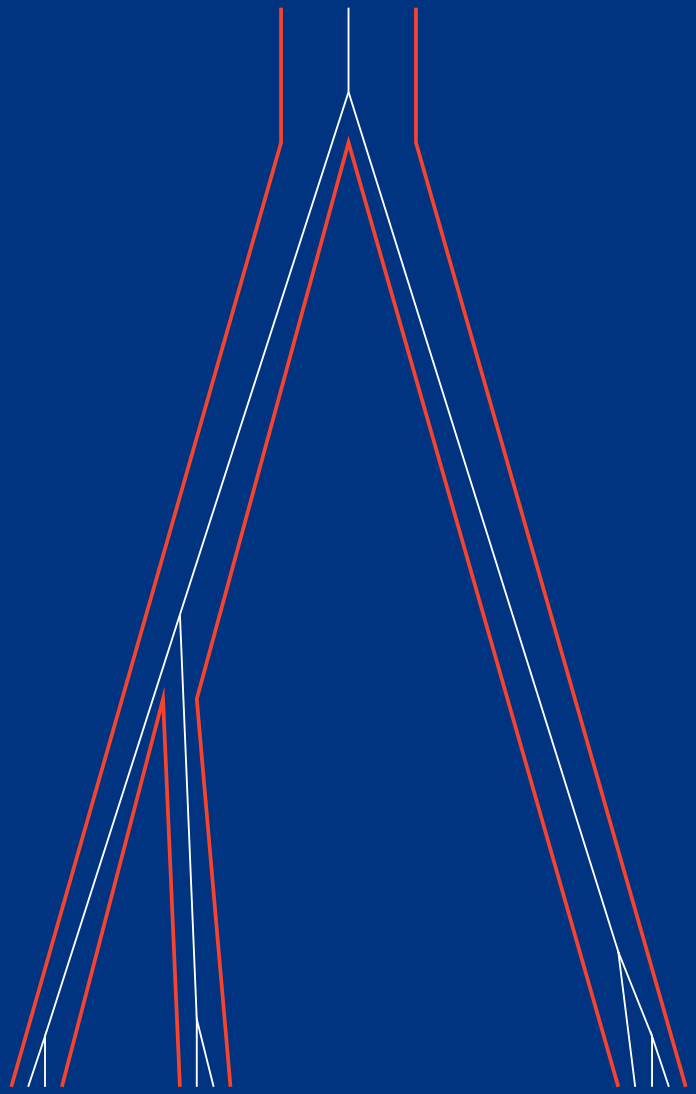
- Even with infinitely many polymorphisms, a substantial fraction of all junctions would not be detected
- In reality, there are clearly too few polymorphisms per junction to estimate the graph
- Remember: a phylogenetic algorithm will *always* reconstruct a tree, regardless of whether there exists a tree to be reconstructed...

We do not in general wish to reconstruct genealogical graphs

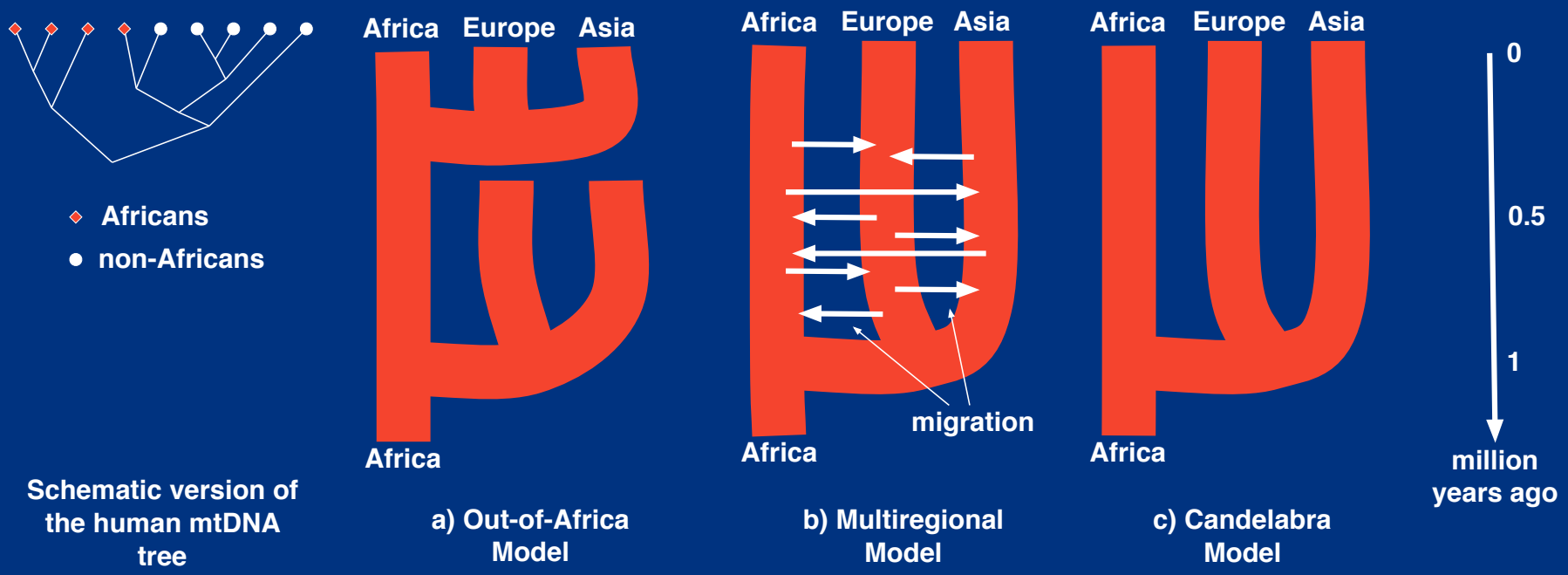
- Population genetics is not phylogenetics!
- Gene genealogies are of no interest *per se* — they are random outcomes of an underlying evolutionary process, and are of interest only insofar as they contain information about this process

Gene trees and species trees

Phylogenetic methods estimate species trees by estimating gene trees; they are appropriate if and only if the latter are strongly correlated with the former



Phylogenetic methods are not applicable to within-species data



- We must consider the likelihood of the data under alternative models

A likelihood framework

Phylogenetics:

$$L = \mathbb{P}(D|G, \mu)$$

Population genetics:

$$L = \sum_{G} \mathbb{P}(D|G, \mu)\mathbb{P}(G, \alpha)$$

Here D is the data, G the genealogy, μ the mutation model, and α the demographic model

Note that G is a *nuisance parameter* in population genetics

Uses of the coalescent

- A mathematical modeling tool
- A simulation tool for hypothesis testing and exploratory data analysis
- The basis for full likelihood inference

The simplicity and elegance of the coalescent process makes it a powerful modeling tool

At least for the standard coalescent, it is often possible to derive results analytically

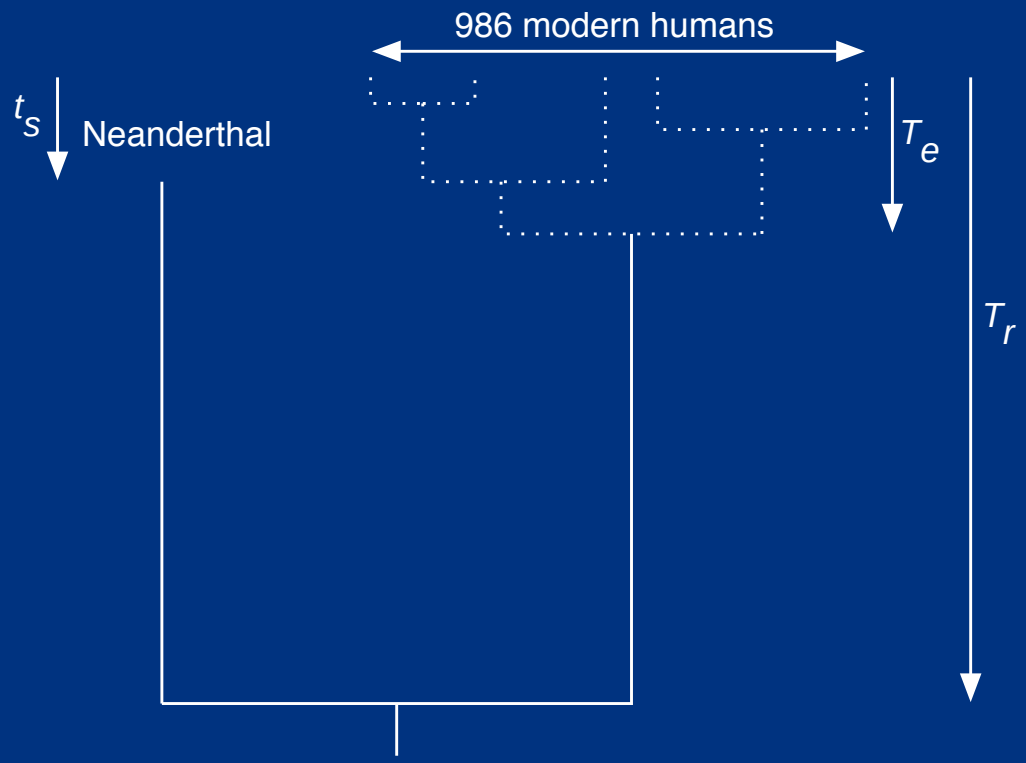
- Estimators and test, *e.g.*, Tajima's D statistic
- Illuminating theoretical results, *e.g.*, the probability that a sample of size n contains the MRCA of the entire population is

$$\frac{n-1}{n+1}$$

Almost any scenario can be simulated using the coalescent

- Coalescent simulations are enormously more efficient than classical methods
- Simulated data can be compared with real data — or used to evaluate the feasibility of a study before it is carried out

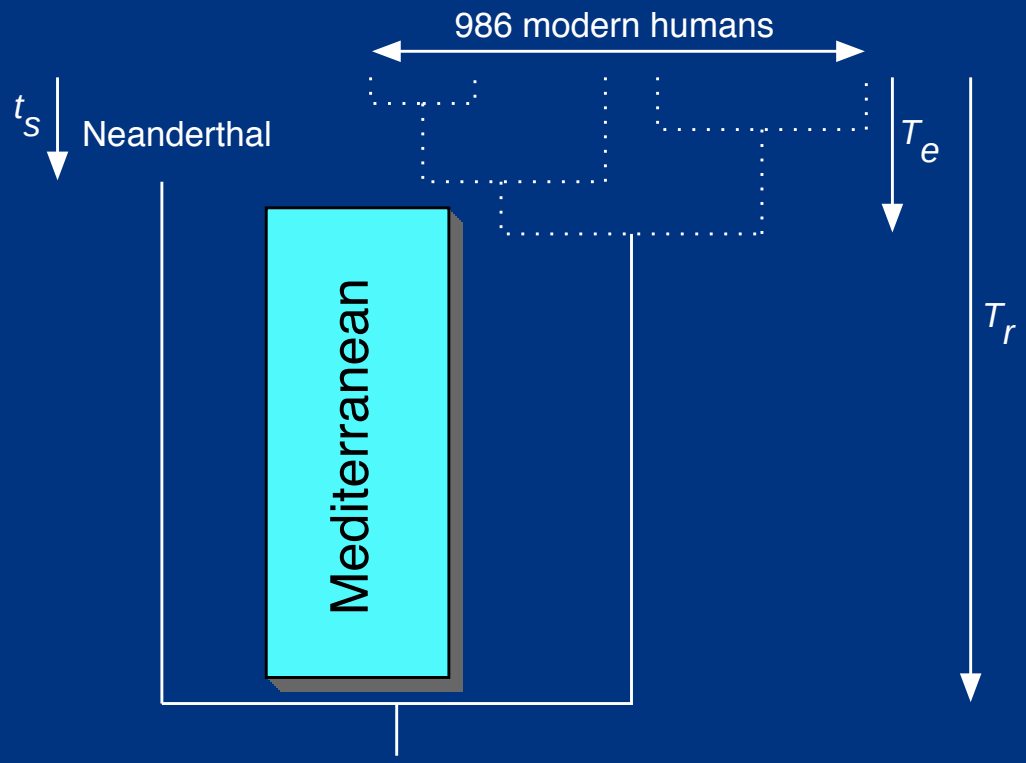
Example: ancient Neanderthal mtDNA



- Modern humans monophyletic
- $T_r > 4T_e$

Does this prove that Neanderthals and modern humans did not interbreed?

Example: ancient Neanderthal mtDNA



Assuming that they did interbreed, what is the probability of getting a tree like the one observed just by chance?

Coalescent simulations showed that this probability is high even for large amounts of interbreeding

Full likelihood analysis

- In principle possible
- In practice difficult
- Unless major breakthroughs are made, not likely to be applicable to genomic polymorphism data

What is the main insight from coalescent theory?

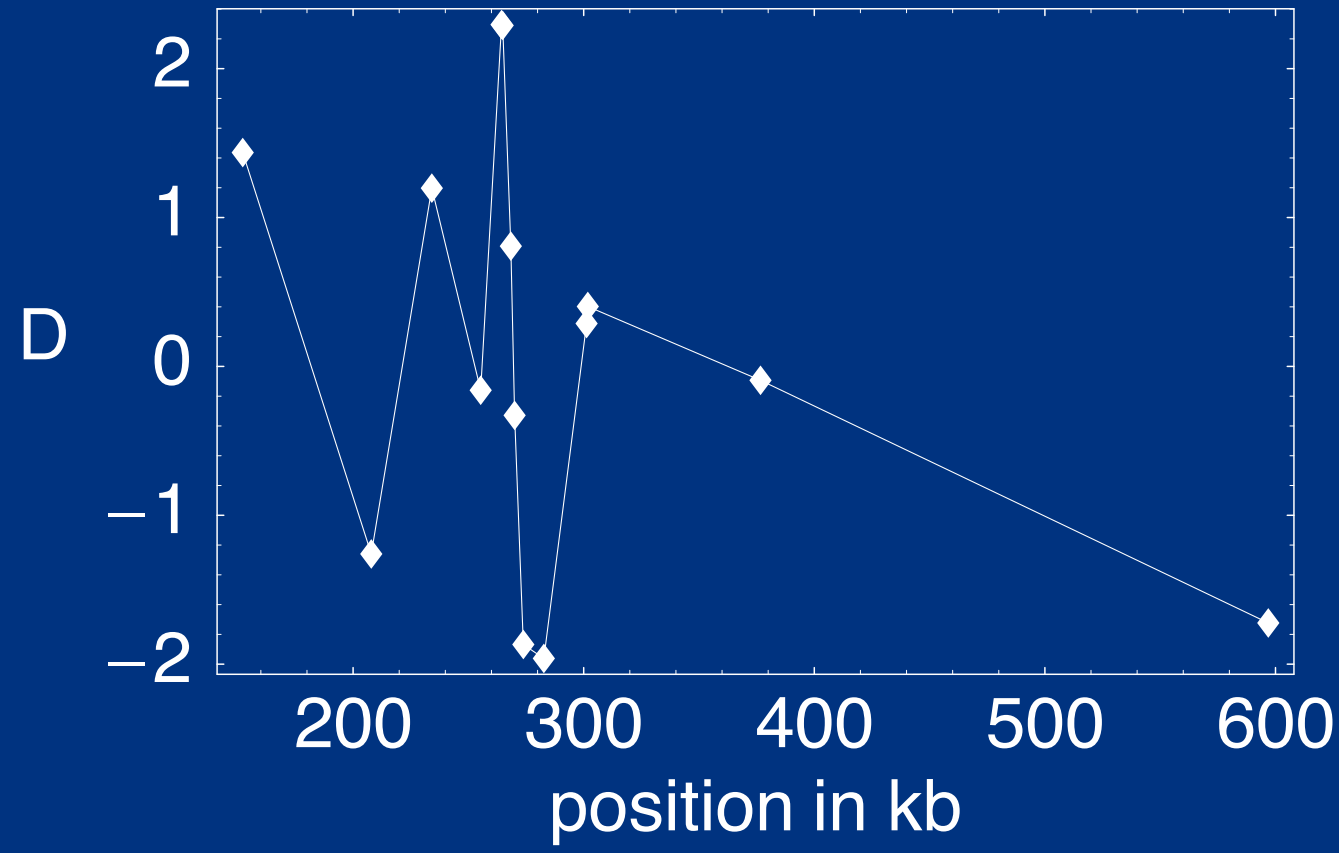
That very large numbers of loci are required to answer most questions!

Population Genomics is upon us!

- Data sets containing 100's and 1000's of loci already exist
- Within 10 years, it seems likely that whole-genome comparisons between species will be common, and that we will have whole genome sequences from 1000's of humans

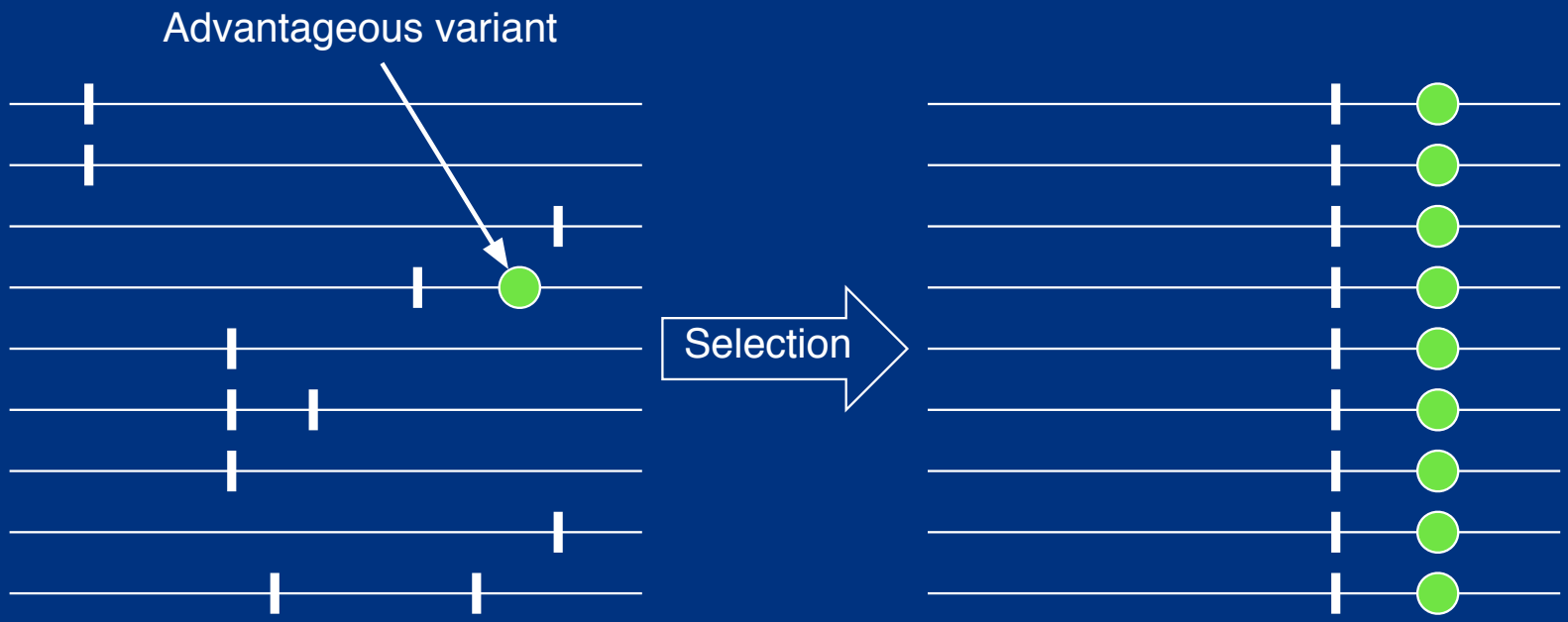
Less assumptions — more data

We will be able to use empirically estimated distributions of test statistics rather than theoretically predicted ones



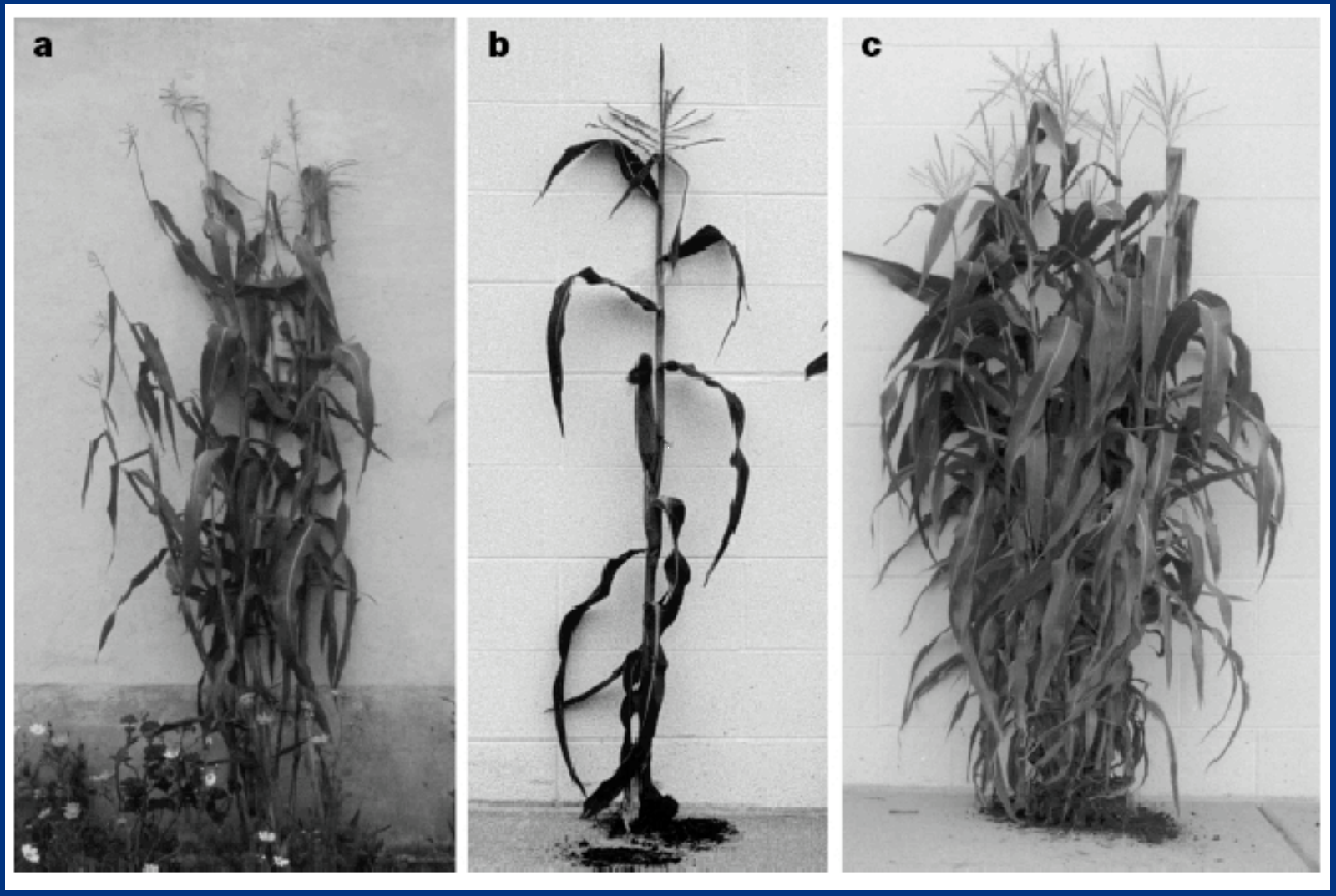
Selective sweeps

- Fixation of new alleles leaves a footprint in the pattern of genomic variation
- Can we find the genes that "make us human"



How many genes?

Teosinte to corn: $<$ 10,000 years; five genes?



teosinte maize maize with *tb1* mutation

What's the use polymorphism data?

- Whole-genome properties
  - demographic (*sensu lato*) history
  - molecular evolution
  - genetic mechanisms
- The history of individual loci — selection
  - divergence between human and other primates
  - traces of selection within the last million years

The history and future of multi-locus methods